Contemporary Engineering Sciences, Vol. 9, 2016, no. 6, 257 - 278
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2016.57208

Goodness-of-Fit Measures for Identification of Factor Models Employing Arbitrarily Distributed Observed Data

Lev S. Kuravsky, Paul A. Marmalyuk and Anastasia S. Panfilova

Moscow State University of Psychology and Education, Computer Science Faculty 29 Sretenka str., 127051 Moscow, Russia

Abstract

The present paper considers further development of factor model analysis intended for monitoring of factors responsible for behavior of technical and other systems. It is presented a new technique of goodness-of-fit measure estimation for the case of unrestricted factor models employing arbitrarily distributed observed data, which is based on the capabilities of self-organizing feature maps (Kohonen networks) and the Monte Carlo method. Obtained results make it possible to avoid undesirable restrictions on observation data inherent in the traditional factor model identification procedure as well as restrictions on model structure used in the previous studies. Following this technique, each observed variance and covariance is associated with an equation that expresses analytically their expected value via free model parameters and equates it with the corresponding sample estimation. The overdetermined set of linear or non-linear equations is usually obtained and, then, solved with the aid of a numerical optimization procedure. Calculation of goodness-of-fit measure is based on comparison of the pseudosolution residual vector and generated random samples of residual vectors corresponding to the solutions being within the pseudosolution neighborhood. Simulated random samples of residual vectors are used to train self-organizing feature maps of proper dimension and, as a result, to obtain samples of Euclidean distances between residual vectors used as input cases and the centers (weight vectors) of "winning" units. That yields the opportunity to calculate the probability of exceeding the distance between the pseudosolution residual vector and its corresponding “winning” unit center and use it as a goodness-of-fit measure. A new measure for obtained pseudosolution precision as well as the

technique for revealing the most probable factor model structure are under consideration.

Keywords: Factor model, goodness-of-fit measure, neural networks, self-organizing feature maps, Monte Carlo method

1 Introduction

Available parameters measured for condition monitoring do not usually represent characteristics of a system under study in the mode that is suitable directly for understanding system status and formulating reliable conclusions sufficient for proper diagnostics. For multivariate measurements, which condition monitoring usually deals with, it is important to reveal some latent factors responsible for joint variability of observed measurable parameters, determine their nature and scope of influences, and use the obtained information to identify system condition.

Therefore, it is desirable to replace the parameters those are easy to measure by the parameters those are easy to interpret and understand the system behavior, with minimal information losses being expected during this data mining. Functional relationships between revealed factors and observed parameters are also to be determined for further analysis. As a result of this study, a researcher should get the structure of causal connections between revealed factors and observed variables as well as immediate factor values to differentiate system status, if necessary.

To meet all the indicated requirements, empirical mathematical models and corresponding methods of multivariate statistical analysis were developed. The most appropriate in the discussed situation are structural equation factor models including exploratory and confirmatory ones as well as methods of their analysis. These approaches are based on the analysis of sample covariance or correlation matrices of the observed parameters under study. The exploratory analysis assumes unknown number of uncorrelated factors with a priori undetermined interpretation, whereas the confirmatory one assumes the factors, their interpretation, causal connections with observed variables and correlation connections between latent factors to be known beforehand. Confirmatory models also admit a convenient technique for estimating statistical significance of each their component. Structural equation models usually specify not only correlation associations between factors but also factor causal dependencies. Thus, structural equation modeling usually includes confirmatory factor analysis of the measurement model part.

Since substantial hypotheses concerning the reasons of possible influences on the observed variables are usually available in practice, the latter approach is preferable.

However, the traditional structural equation modeling has its own intrinsic defects:

− It needs solution of the laborious local multivariate optimization problem to estimate the values of free model parameters that results in impossibility of the global minimum estimation and ambiguous solution

− Multivariate normality of observed variables is necessary to get convenient goodness-of-fit criterion for model identification

− Optimization criterion in use is too exigent in case of relatively large samples of observation data.

To overcome aforesaid problems, new approaches were developed. Their features and advantages were originally presented in papers[3-8]. Presented here is further development of this approach in application to the goodness-of-fit factor model analysis, which is based on the capabilities of self-organizing feature maps and the Monte Carlo method. Obtained results make it possible to avoid undesirable restrictions on observation data inherent in the traditional factor model identification procedure as well as restrictions on model structure used in the previous studies.

2 Structural Equation Modeling

Strictly determined factor model of the phenomenon under study is assumed in the traditional structural equation modeling. A factor model that connects both latent and observed variables is formed using knowledge about the application domain. The hypotheses concerning the model structure have to be based on the analysis of the investigated factors nature. It is admissible to formulate quantitative assumptions concerning correlations between latent variables as well as factor loadings. Free model parameters are calculated to get the best approximation of correlation or covariance matrices for observed variables from the viewpoint of a given criterion.

Objects of the traditional factor analysis are correlation or covariance matrices for observed variables. Purpose of the analysis is to find model parameters that explain variability of observations with acceptable errors.

In using the maximum likelihood method the following function is to be minimized as a criterion for selection of free parameters:

$$\boldsymbol{F = [ln\,|\Sigma| - ln\,|S| + tr(S\Sigma^{-1}) - m]\,(N-1),}$$

where $\mathbf{S}$ – sample covariance matrix for observed variables, $\boldsymbol{\Sigma}$ – expected covariance matrix for observed variables, $|\boldsymbol{\Sigma}|$ and $|\mathbf{S}|$ – determinants of matrices $\boldsymbol{\Sigma}$ and $\mathbf{S}$, $\boldsymbol{tr}\,(\mathbf{S\Sigma^{-1}})$ – trace of matrix $(\mathbf{S\Sigma^{-1}})$, $\boldsymbol{N}$ – size of the sample used to calculate matrix $\mathbf{S}$, $\boldsymbol{m}$ – number of observed variables.

Elements of the expected covariance matrix are analytical expressions consisting of free model parameters. In case of multivariate normality of observed variables values of the criterion F are distributed as χ^2.

3 Estimating goodness-of-fit measures with the aid of self-organizing feature maps

To estimate the values of free model parameters in case of the above-stated minimization criterion it is necessary to solve numerically a laborious local multivariate optimization problem by the acceptable iteration methods. In general case, this way results in impossibility of the global minimum estimation, since one of the possible local minima depending on its initial approximation is usually found. Consequently, the solution is ambiguous.

Suggested alternative variant of the confirmatory factor analysis[3-4] allows to find the values of free model parameters by a direct (noniterative) method ensuring an unambiguous optimal solution.

In the alternative variant of the confirmatory factor analysis one has to:

- Compose an overdetermined set of the equations each of which expresses observed variances and covariances via free factor variances and covariances with the aid of a factor model
- Solve them by a direct (noniterative) method using a certain form of the maximum likelihood approach, which is different from the one used in the confirmatory factor analysis[3]
- Examine for the adequacy of the obtained equation sets to observations with the aid of statistical goodness-of-fit tests.

To avoid solving non-linear equation sets as respects to free correlation coefficients and factor loadings the variance components model[9] in which path coefficients (factor loadings) equal to unity is in use.
Hereinafter, each observed variance and covariance is associated with an equation that expresses analytically their expected value via free variances and covariances of latent variables and equates it with the corresponding sample estimation. The set of the equations is obtained, in which number of the equations equals to the number of observed variances and covariances. If this number of equations exceeds the number of free model parameters, the overdetermined set of equations is the case. It is last situation that is necessary for the further decision. The method under consideration needs also multivariate normality of observed variables.
Provided that the variance components path model is used, the obtained overdetermined set of equations can be represented in matrix notation:

$$\mathbf{Ax=b},$$

where $\mathbf{A}$ - system nxm matrix, which coefficients are determined using the factor model (path diagram) under consideration; **b** - column nx1 vector of variance

and covariance sample estimates, which are determined using observation results; $\mathbf{x}$ - column $\boldsymbol{m}\mathrm{x}\boldsymbol{1}$ vector of unknown free model parameters of interest (viz.: variances and covariances for latent variables). The vector $\boldsymbol{\varepsilon}=\mathbf{A}\mathbf{x}_*-\mathbf{b}$ represents residual of the given set pseudosolution $\mathbf{x}_*=(\mathbf{A}^T\mathbf{V}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{V}^{-1}\mathbf{b}$ obtained by the least-squares method. Assuming in the general case that components of the residual vector are correlated let us express their nonsingular covariance matrix as $\boldsymbol{\sigma}^2\mathbf{V}$.

If

- The equation set matrix is nonsingular (*rank* $\mathbf{A}=\boldsymbol{m}$)
- The transformed residual vector $\mathbf{V}^{-½}\boldsymbol{\varepsilon}$[1] has multivariate normal distribution
- $\mathbf{x}_*=(\mathbf{A}^T\mathbf{V}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{V}^{-1}\mathbf{b}$ is pseudosolution,

then this pseudosolution is a maximum likelihood estimate and statistics

$$\boldsymbol{X}^2=(\mathbf{b}-\mathbf{A}\,\mathbf{x}_*)^T\mathbf{V}^{-1}(\mathbf{b}-\mathbf{A}\,\mathbf{x}_*)/\sigma^2$$

has $\boldsymbol{\chi}^2$-distribution with $\boldsymbol{n-m}$ degrees of freedom.

Last statistics makes it possible to evaluate the model validity level. Under the assumptions indicated above, the presented statistics $\mathbf{X}^2$ makes it possible to test the hypothesis of representability of sample variances and covariances constituting the vector $\mathbf{b}$ with the aid of variances and covariances of latent variables contained in the model under study. Acceptance region is $\mathbf{X}^2\leq\boldsymbol{\chi}^2_{\mathbf{n-m;\alpha}}$ where $\boldsymbol{\alpha}$ is criterion significance level.

As in the traditional confirmatory factor analysis, the considered method also allows making conclusions on statistical significance of different model components under study using goodness-of-fit tests.

To do this one should compare $\boldsymbol{X}^2$ statistics for two models: saturated model containing the component of interest and simplified model where this component is absent (equals to zero.) Let's denote hypothesis that the saturated model coincides with observation results as $\boldsymbol{H_f}$. Significance level of the component of interest is revealed if there is no grounds to discard hypothesis $\boldsymbol{H_f}$. At first one

---

[1] Where $\mathbf{V}=\mathbf{V}^{½}\mathbf{V}^{½}$. The only symmetric nonnegatively defined matrix $\mathbf{V}^{1/2}$, which is called the square root of $\mathbf{V}$, exists for every symmetric nonnegatively defined covariance matrix $\mathbf{V}$, so that $(\mathbf{V}^{1/2})^2=\mathbf{V}$.

should estimate free parameters of the simplified model. The obtained value for X^2 statistics is compared with similar characteristics for the saturated model.
Since the difference in these statistics is asymptotically distributed as χ^2 with the number of degrees of freedom that is equal to the difference in degrees of freedom of saturated and simplified models, this difference is used to verify the zero hypothesis H_r that the simplified model coincides with the observation results against alternative hypothesis H_f.

If H_r hypothesis is not discarded at the given significance level then the component under study is treated as statistically insignificant and the conclusion is made that the available data do not evidence the influence of the studied model part on the observed characteristic under consideration. If H_r hypothesis is discarded (and H_f hypothesis is accepted), then one can talk about the influence of the studied component on the given characteristic.

Advantages of the suggested technique are:

- The problem solution is not reduced to the local multivariate optimization
- Since this method is direct there is no multiplicity of solutions
- No need in search of global minima.

Correct usage of the maximum likelihood method criteria mentioned above for both the traditional and alternative confirmatory factor analysis to identify the values of free model parameters and estimate the model goodness-of-fit measure needs testing multivariate normality of distributions of either observed variables or residual vector components. This procedure is laborious and frequently impossible because of deficiency in observed data[2].

To overcome this problem a new technique that combines the capabilities of self-organizing feature maps (SOFM)[2], or Kohonen networks, and the Monte Carlo method was proposed[4-5] for the variance components factor models. Its framework is presented in Figure 1.

[2] The maximum likelihood criteria in case of the traditional confirmatory factor analysis is also too sensitive to a sample size: small deviations from expected characteristics result in considerable goodness-of-fit measures for large samples.

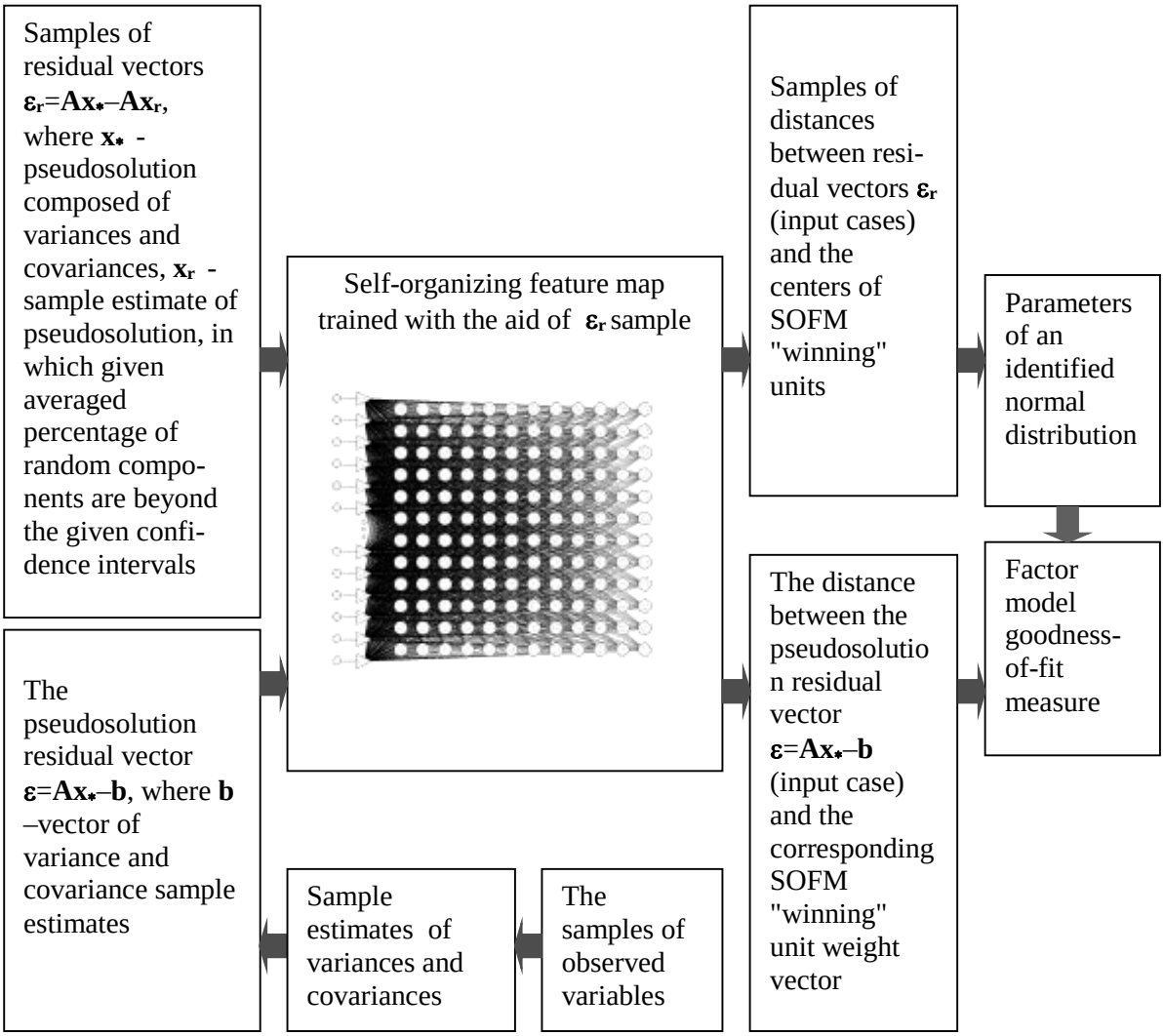

Figure 1. Calculation of variance components factor model goodness-of-fit measure with the aid of the self-organizing feature maps and the Monte Carlo method.

Calculation of goodness-of-fit measures is based on comparison of the pseudosolution residual vector $\boldsymbol{\varepsilon}$ and random samples of residual vectors $\boldsymbol{\varepsilon}_r=\mathbf{A}\mathbf{x}_*-\mathbf{A}\mathbf{x}_r$, where $\mathbf{x}_r$ was pseudosolution estimate composed of variances and covariances, in which given averaged percentage of random components are beyond the given confidence intervals. Residual vectors $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}_r$ were both resulted from a factor model under consideration

Random samples of residual vectors $\boldsymbol{\varepsilon}_r$ are used to train[3] self-organizing feature maps of proper dimension and, as a result, to obtain samples of Euclidean distances between residual vectors $\boldsymbol{\varepsilon}_r$ used as input cases and the centers (weight vectors) of SOFM "winning" units. Taking into account the structure of the Euclidian distance and high dimension of residual vectors, which is typical for practical applications, these samples of distances are to be about normally distributed. Estimation of their means and variances identified the given distributions and yielded the opportunity to calculate the probability of exceeding the distance between the pseudosolution residual vector $\boldsymbol{\varepsilon}$ and its corresponding "winning" unit center. This probability is considered as a factor model goodness-of-fit measure.

To get information about possible deviations of identified parameters from their estimations obtained with the aid of a given factor model, series of samples with given changing averaged percentages[4] of random components going beyond the given confidence intervals were generated for SOFM training. Comparison of above-stated distance distributions for different percentages makes it possible to reveal the maximum likelihood component-wise structure of significant deviations for the pseudosolution components.

In the presented procedure SOFM ensures clustering residual vectors which connect point $\mathbf{A}\mathbf{x}_*$ and points $\mathbf{A}\mathbf{x}_r$, with centers of radial basis function units representing obtained cluster centers. Dimensions of the corresponding topological map are extended up to the limits when neither maximum unit win frequency diminishes nor number of winning units increases for the generated $\mathbf{x}_r$ sample. In this case distances between a residual vectors $\boldsymbol{\varepsilon}_r$ and their winning cluster centers do not represent geometric loci of points $\mathbf{A}\mathbf{x}_r$ from the SOFM point of view. Therefore, they can be considered to be under the SOFM threshold of sensitivity and, so, should represent permissible non-significant variations ("noise") of residual distances from the nearest cluster center, which are not conditioned by geometric causes.

---

[3] SOFM are typically arranged with the two-dimensional layer (topological map) consisting of radial basis function units. From an initially random set of radial unit centers, a training algorithm tests each input case and selects the nearest radial or "winning" unit. The center of this unit and its neighbors' centers are then modified iteratively to be more like the input case.

[4] From 0% to 100% by certain intervals.

Geometric illustration clarifying the above-stated procedure is given in Figure 2[5].

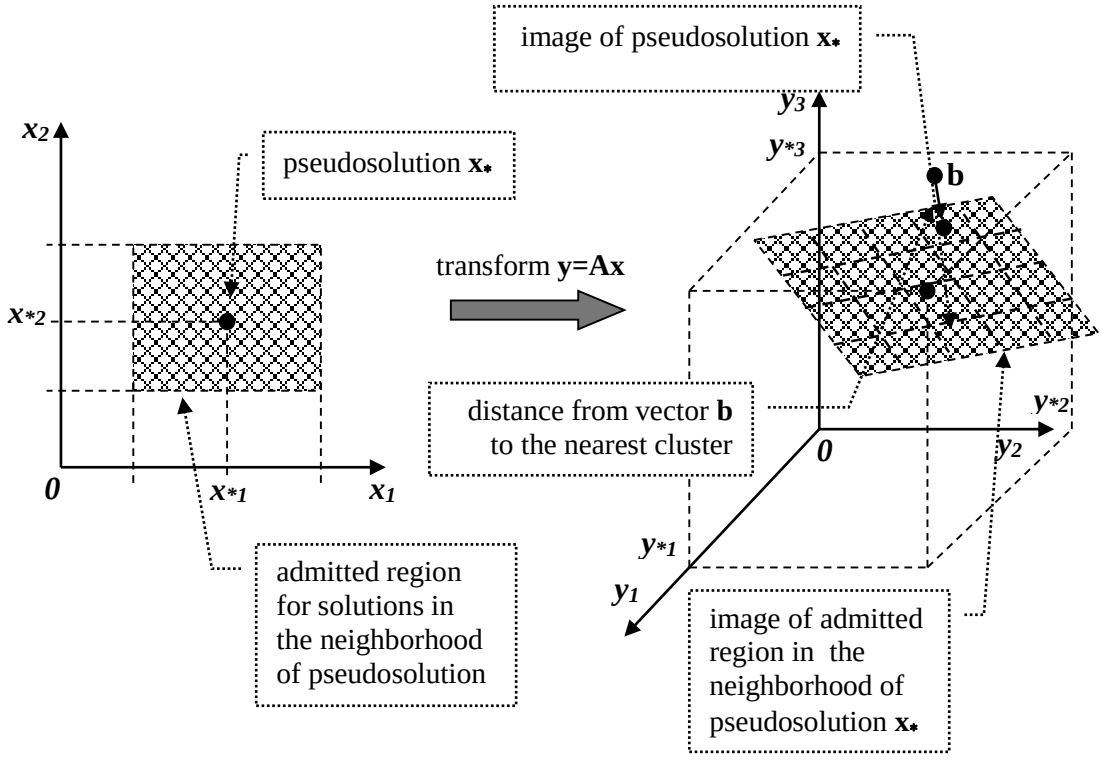

Figure 2. Geometric interpretation of the admitted solution region linear transform.

The technique used for sample generation of residual vectors $\boldsymbol{\varepsilon_r}=\mathbf{Ax_*}-\mathbf{Ax_r}$ is shown in Figure 3. According to this approach identified model variances and covariances which compose the pseudosolution were then repeatedly converted to sets of simulated sample estimates of corresponding variances and correlations. Generation of the given test samples makes it possible to consider the presented technique as a form of the Monte Carlo method.

[5] According to the singular decomposition theorem for linear transform of some vector space to the vector space of greater dimension, in general case, image of ***n***-dimensional parallelepiped in the space of some greater dimension is a transformed ***n***-dimensional parallelepiped.

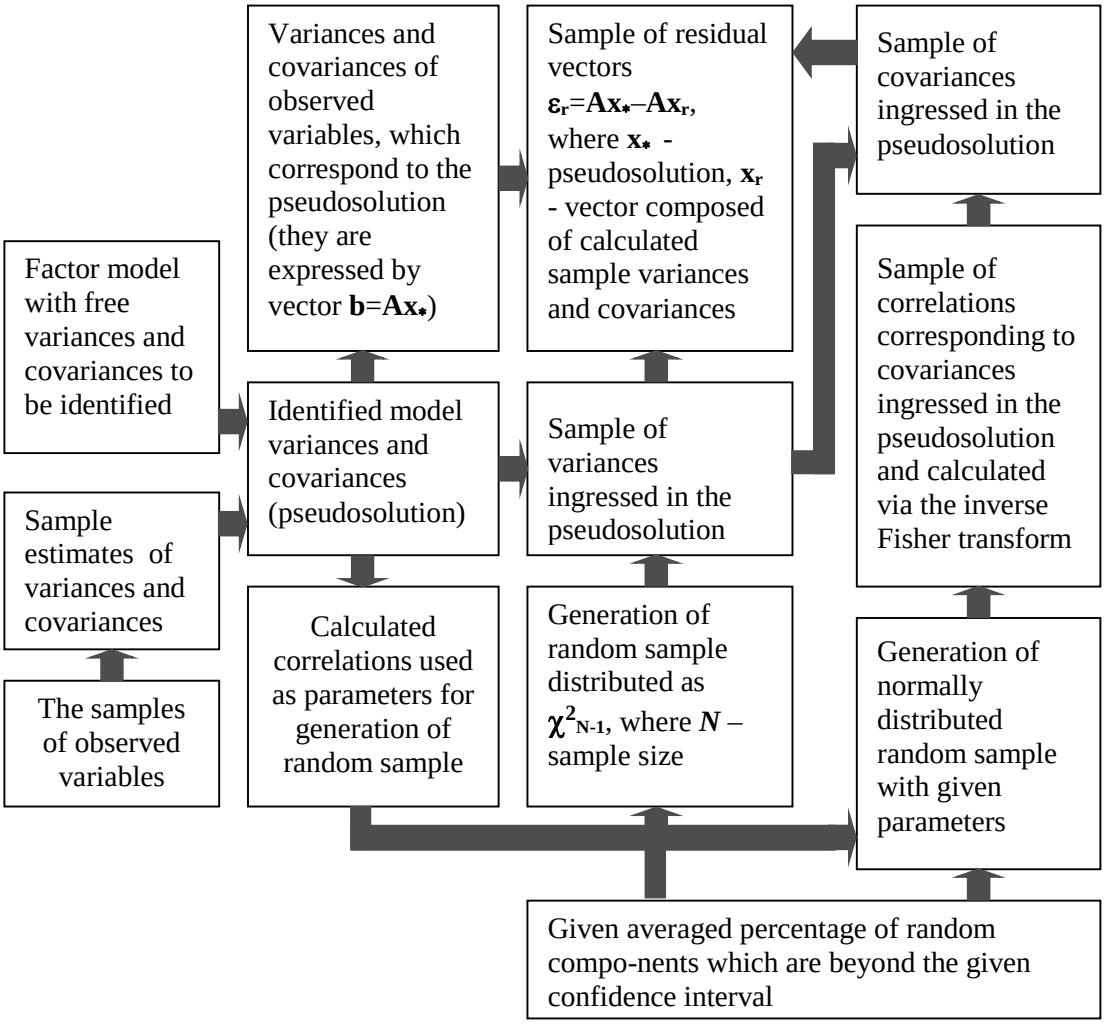

Figure 3. Generation of a sample of residual vectors $\varepsilon_r=Ax_*-Ax_r$, where x_* - pseudosolution, x_r - vector composed of calculated sample variances and covariances.

Sample estimates of variances are calculated using the following formula derived from the expression for distribution of sample variance of normally distributed random variable[6]:

$$V_S = \frac{\chi^2_{N-1} V}{N-1},$$

where V_S is variance sample estimate, N is sample size specified for generation, V is an identified variance ingressed in the pseudosolution, χ^2_{N-1} is a random

[6] Thus, sample estimates of corresponding pseudosolution components are considered as normally distributed ones. This assumption does not yield any restrictions on observation data.

element distributed as χ^2 with $N-1$ degrees of freedom. Elements χ^2_{N-1} are software generated.

Sample estimates of covariances are calculated via corresponding sample estimates of correlations using the fact of approximate distribution normality for their Fisher transform, viz.: distribution of the statistics[7]

$$z = \frac{1}{2} \ln \frac{1+r}{1-r},$$

where r is sample correlation, can be approximated by the normal distribution with the expectation

$$z = \frac{1}{2} \ln \frac{1+\rho}{1-\rho},$$

where ρ is correlation value, and the variance

$$\frac{1}{N-3}.$$

Samples of Fisher transform results are software generated for each covariance ingressed in the pseudosolution, with correlations ρ being substituted for corresponding correlations. Required correlations themselves are restored by means of calculating the inverse Fisher transform for the abovementioned generated values. After that they are converted into the covariances ingressed in the pseudosolution.

Simulated samples of variances and covariances yield required samples of residual vectors $\boldsymbol{\varepsilon}_r = \mathbf{A}\mathbf{x}_* - \mathbf{A}\mathbf{x}_r$ used for SOFM training.

Some principal advantages of the suggested alternative variant of the confirmatory factor analysis were caused by linearity of its problem formulation resulted from use of the variance components models. However, many application problems of interest cannot be solved within the framework of such standard.

Presented in this paper is further development of the above-stated factor approach, in which arbitrary structural equation models without reciprocal variable causations creating "feedback-loops" are acceptable[8]. Extension of the model set results in loss of obtained equation set linearity and, therefore, in loss of possibility for identifying free model parameters by a simple direct procedure.

Following the suggested technique one has to:

---

[7] That defines the Fisher transform.

[8] These models include path coefficients ones.

- Express observed variances and covariances via free parameters of a factor model in use to compose the overdetermined set of equations which are non-linear in general case
- Calculate the pseudosolution of this set of equations by some numerical optimization procedure
- Examine for the adequacy of the obtained pseudosolution to observations with the aid of some acceptable statistical goodness-of-fit test.

The mentioned overdetermined set of equations can be expressed the following way:

$$\mathbf{F(x)=b},$$

where $\mathbf{F(x)}$ - $\boldsymbol{n}$-dimensional non-linear operator applied to $\boldsymbol{m}$-dimensional vector $\mathbf{x}$ of unknown free model parameters of interest, which $\boldsymbol{n}$ components are expected analytic expressions of variances and covariances for observed variables via $\boldsymbol{m}$ free parameters of a factor model under consideration; $\mathbf{b}$ - column $\boldsymbol{n}$x$\boldsymbol{1}$ vector of variance and covariance sample estimates, which are determined using observation results.

The vector $\boldsymbol{\varepsilon}=\mathbf{F(x_*)-b}$ represents a residual of the pseudosolution $\mathbf{x_*}$, where

$$\| \mathbf{F(x_*)} - \mathbf{b} \| = \min_{\mathbf{x} \in \mathbf{X}} \| \mathbf{F(x)} - \mathbf{b} \|,$$

$\|.\|$ - vector Euclidean norm, $\mathbf{X}$ – admitted region for $\mathbf{x}$.

To get this pseudosolution, any available numerical non-linear multivariate local optimization procedure with a minimization criterion represented by the residual Euclidean norm can be used. Gradient techniques are acceptable for this purpose. In particular, the authors employed a procedure called the Generalized Reduced Gradient.

To examine for the adequacy of the calculated pseudosolution to observations further development of the above-stated technique based on both the SOFM capabilities and the Monte Carlo method is suggested here. Its framework is shown in Figure 4. As before, calculation of goodness-of-fit measure is based on comparison of the pseudosolution residual vector $\boldsymbol{\varepsilon}=\mathbf{F(x_*)-b}$ and random samples of residual vectors $\boldsymbol{\varepsilon_r}= \mathbf{F(x_*)-F(x_r)}$, where $\mathbf{x_r}$ is a generated random test vector belonging to a given neighborhood of the pseudosolution $\mathbf{x_*}$.

Any arbitrary distribution may be assigned to vectors $\mathbf{x_r}$, nevertheless for practical purposes it is convenient to produce them normally distributed, with the standard deviation being varied. If necessary, given averaged percentage of random vector components are placed beyond the given neighborhood intervals. Random samples of residual vectors $\boldsymbol{\varepsilon_r}$ are used to train SOFM of proper dimension and, as a result, to obtain samples of Euclidean distances between residual vectors $\boldsymbol{\varepsilon_r}$ used as network input cases and the centers of SOFM "winning" units. These samples

are close to normally distributed ones owing to the afore-cited reasons. Estimation of their means and variances identifies the given distributions and yields the opportunity to calculate the probability of exceeding the distance between the pseudosolution residual vector $\boldsymbol{\varepsilon}$ and its corresponding "winning" unit center. This probability is considered as a factor model goodness-of-fit measure.

To get information about possible deviations of identified parameters from their estimations obtained with the aid of a given factor model, series of samples with both given different standard deviations of random components and their changing averaged percentages of going beyond the given neighborhood intervals are generated for SOFM training. For ease of analysis standard deviation of each test vector component is assumed to be equal to a certain constant percentage of the corresponding component mean value. Comparison of the above-stated SOFM distance distributions for different standard deviations and percentages makes it possible to reveal the maximum likelihood combination of the obtained pseudosolution precision presented by the estimated standard deviation and the component-wise structure of significant deviations for the pseudosolution components.

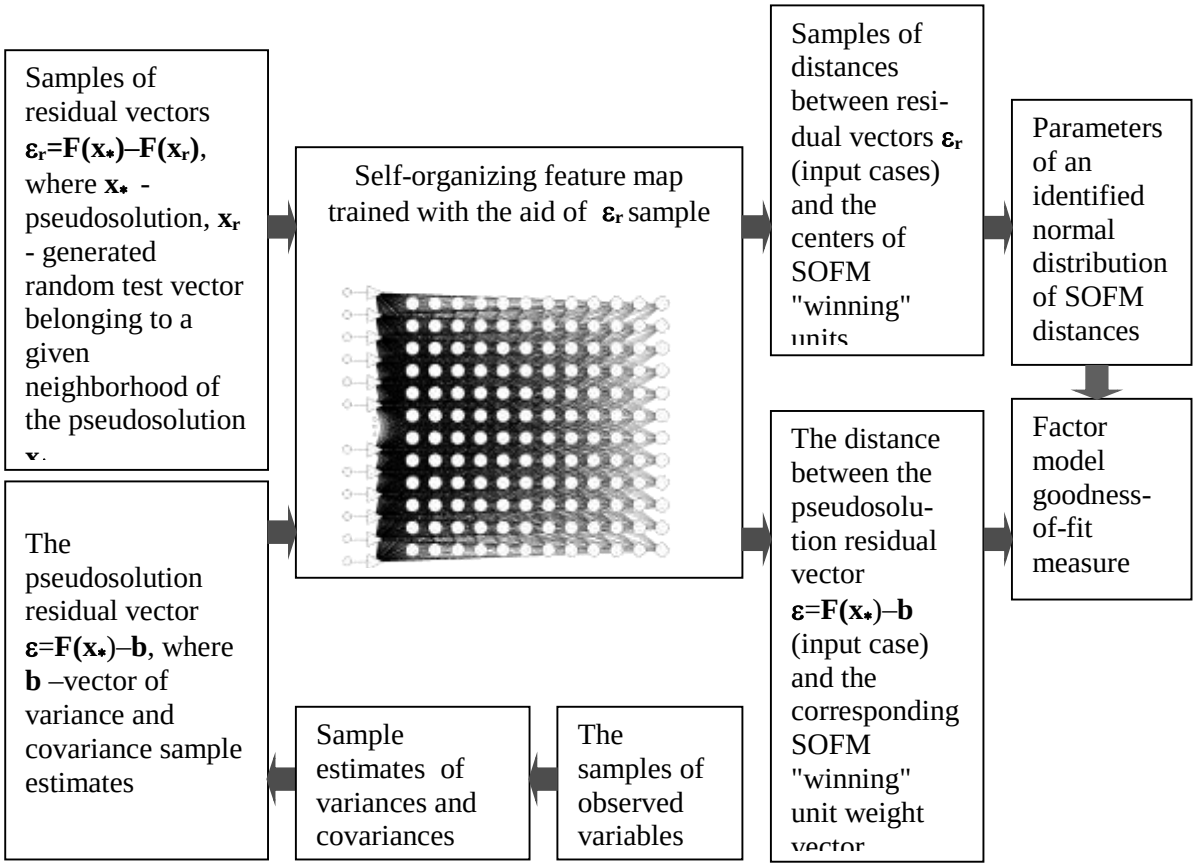

Figure 4. Calculation of arbitrary factor model goodness-of-fit measure with the aid of the self-organizing feature maps and the Monte Carlo method.

Geometric illustration clarifying the above-stated procedure, which is non-linear in general case, is given in Figure 5.

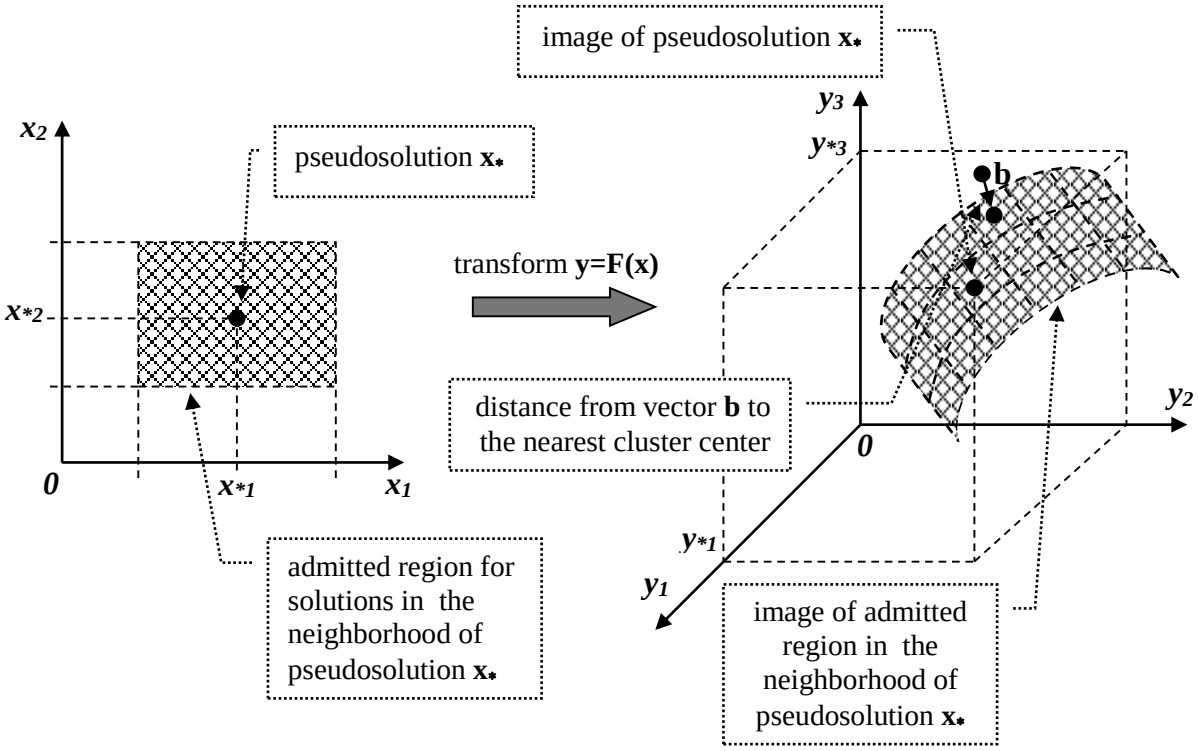

Figure 5. Geometric interpretation of the admitted solution region non-linear transform.

The suggested approach allows making conclusions on statistical significance of differences between two most probable factor model patterns under study using certain probability tests. Specific parameters of these model patterns can be identified by the foregoing technique. To compare patterns one should consider their maximum likelihood ratios $r=\sigma/m$, where σ is the most probable standard deviation for generated normally distributed values of free model parameters and m – corresponding distribution mean value. Since standard deviations of these generated values are assumed here to be equal to a certain constant percentage of relevant mean values, these ratios are kept constant for all model pattern parameters, but can differ for various patterns which allow, in general case, diverse averaged percentages going beyond the given parameter neighborhood intervals.

Let the ratios of compared patterns equal to $r_1=\sigma_1/m_1$ and $r_2=\sigma_2/m_2$, correspondingly, and $r_1 \le r_2$. Comparison is carried out for the same relative stan-

dard deviation $\sigma_* = r_1 = r_2 m_2$ when the mean value m_1 equals to 1. In this case probability of the obtained deviation of reduced mean $m_2 = r_1 / r_2$ is estimated, viz.: probability $P(m_2 \le X \le 1) = \Phi(1) - \Phi(m_2)$ of being within the limits $[m_2; m_1 = 1]$ is calculated for random quantity X, where Φ is the normal distribution function with a mean of 1 and a standard deviation of σ_*. If this probability is greater than the given significance level that is usually equal to 0.05, the pattern difference is recognized as significant, otherwise it is considered as negligible.

The goodness-of-fit measures under consideration give the opportunity to determine the sample sizes required for testing hypotheses of equality of the distance between the pseudosolution residual vector ε and its corresponding SOFM "winning" unit center to the certain value with both the given significance level and given test power. A formula of interest is derived from the comparison of corresponding acceptance region limits[1]:

$$N = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{d_{norm}} \right)^2,$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are standard normal distribution quantiles of orders $1-\alpha/2$ and $1-\beta$, correspondingly; α is significance level; β is probability of type 2 error; d_{norm} is the ratio of deflection of true distance expectation from the tested certain value to the standard deviation of distance distribution.

Advantages of the presented techniques for estimating goodness-of-fit measures are the following:

- Arbitrary structural equation factor models without reciprocal
- causal relationships of variables leading to "feedback-loops" are acceptable
- No need to test multivariate normality of distributions of either observed variables or residual vector components
- It is possible to reveal the maximum likelihood combination of the obtained pseudosolution precision and the component-wise structure of significant deviations for the pseudosolution components
- Simple procedure of estimating type 2 statistical errors is available
- Higher reliability of obtained goodness-of-fit measures because of unrestrictedness of generated random samples of the pseudosolution components and the following unlimited goodness-of-fit estimation accuracy.

The presented techniques were software implemented on the base of the *LabVIEW* graphical programming environment. The work of self-organizing feature maps was simulated with the aid of the STATISTICA Neural Networks software package.

4 Illustrative example

To illustrate the presented approach to estimation of goodness-of-fit measures, results of its application for studying psychometric intelligence are shown here. Under consideration were two intelligence factor models[10] presented in Figures 6 and 7.

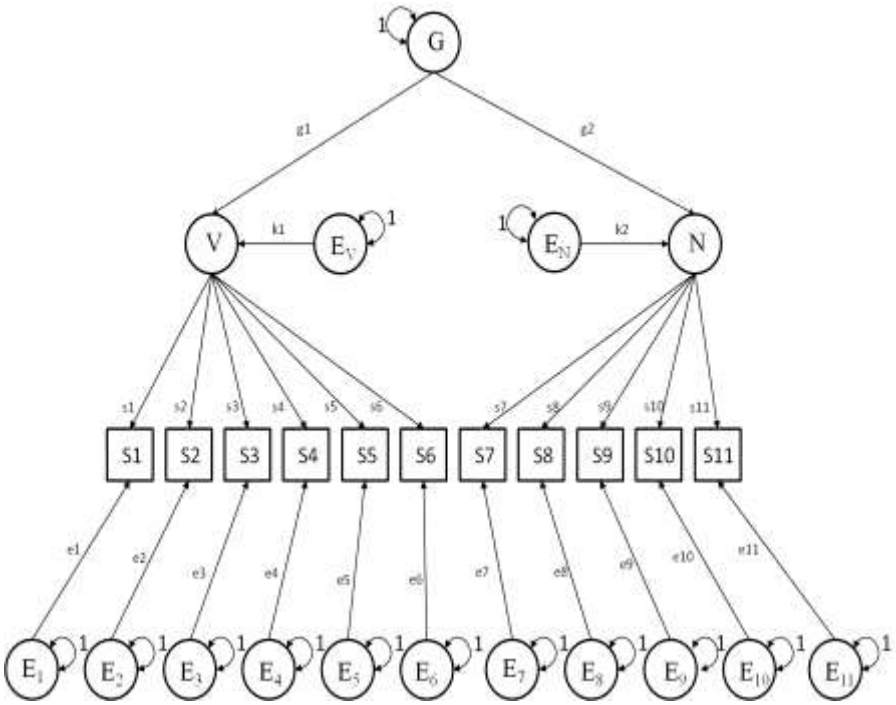

Figure 6. Hierarchical intelligence factor model (G, V, N, E_V, E_N, E_1, E_2, …, E_{11} – latent factors; S_1, S_2, …, S_{11} – observed variables).

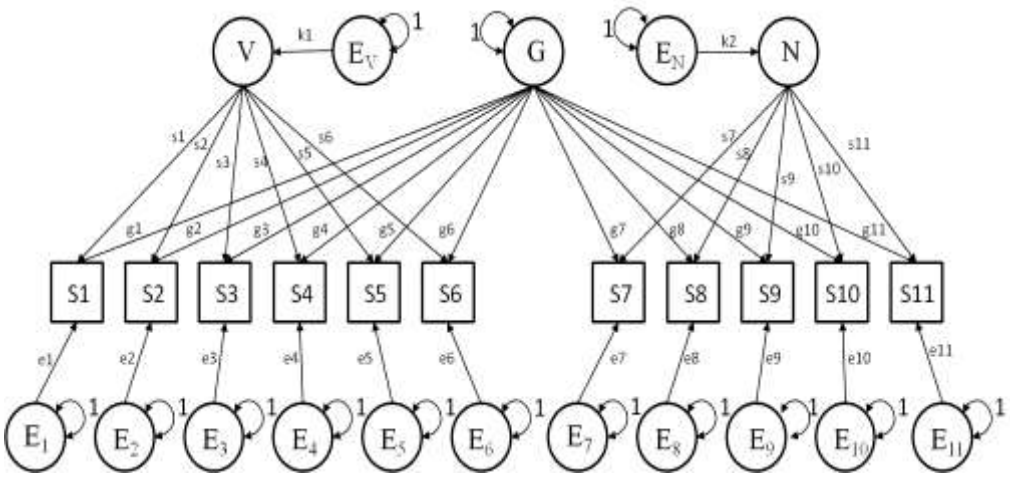

Figure 7. Nested intelligence factor model (G, V, N, E_V, E_N, E_1, E_2, …, E_{11} – latent factors; S_1, S_2, …, S_{11} – observed variables).

Expected covariance matrices for these models are given in Tables 1 and 2. Each of the shown analytical matrix elements after equating it to a corresponding covariance or variance sample estimate for relevant observed parameters selected from the set $\{S_1, S_2, \ldots, S_{11}\}$ yields a non-linear equation for n-dimensional non-linear operator $\mathbf{F(x)}$ applied to m-dimensional vector of unknown free model parameters $g_1, g_2, \ldots, g_{11}, k_1, k_2, s_1, s_2, \ldots, s_{11}, e_1, e_2, \ldots, e_{11}$, where $n=66$, $m=26$ for the first model but $m=35$ for the second one (see Section 3).

Table 1. Expected covariance matrix for the hierarchical factor model.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
S1	$s_1^2(g_1^2+k_1^2)+e_1^2$										
S2	$s_1s_2(g_1^2+k_1^2)$	$s_2^2(g_1^2+k_1^2)+e_2^2$									
S3	$s_1s_3(g_1^2+k_1^2)$	$s_2s_3(g_1^2+k_1^2)$	$s_3^2(g_1^2+k_1^2)+e_3^2$								
S4	$s_1s_4(g_1^2+k_1^2)$	$s_2s_4(g_1^2+k_1^2)$	$s_3s_4(g_1^2+k_1^2)$	$s_4^2(g_1^2+k_1^2)+e_4^2$							
S5	$s_1s_5(g_1^2+k_1^2)$	$s_2s_5(g_1^2+k_1^2)$	$s_3s_5(g_1^2+k_1^2)$	$s_4s_5(g_1^2+k_1^2)$	$s_5^2(g_1^2+k_1^2)+e_5^2$						
S6	$s_1s_6(g_1^2+k_1^2)$	$s_2s_6(g_1^2+k_1^2)$	$s_3s_6(g_1^2+k_1^2)$	$s_4s_6(g_1^2+k_1^2)$	$s_5s_6(g_1^2+k_1^2)$	$s_6^2(g_1^2+k_1^2)+e_6^2$					
S7	$s_1s_7 g_1g_2$	$s_2s_7 g_1g_2$	$s_3s_7 g_1g_2$	$s_4s_7 g_1g_2$	$s_5s_7 g_1g_2$	$s_6s_7g_1g_2$	$s_7^2(g_2^2+k_2^2)+e_7^2$				
S8	$s_1s_8 g_1g_2$	$s_2s_8 g_1g_2$	$s_3s_8 g_1g_2$	$s_4s_8 g_1g_2$	$s_5s_8 g_1g_2$	$s_6s_8g_1g_2$	$s_7s_8(g_2^2+k_2^2)$	$s_8^2(g_2^2+k_2^2)+e_8^2$			

Table 1. (Continued): Expected covariance matrix for the hierarchical factor model.

S9	$s_1s_9 g_1g_2$	$s_2s_9 g_1g_2$	$s_3s_9 g_1g_2$	$s_4s_9 g_1g_2$	$s_5s_9 g_1g_2$	$s_6s_9g_1g_2$	$s_7s_9 (g_2^2+k_2^2)$	$s_8s_9 (g_2^2+k_2^2)$	$s_9^2(g_2^2k_2^2)+e_9^2$		
S10	$s_1s_{10} g_1g_2$	$s_2s_{10} g_1g_2$	$s_3s_{10}g_1g_2$	$s_4s_{10}g_1g_2$	$s_5s_{10} g_1g_2$	$s_6s_{10} g_1g_2$	$s_7s_{10} (g_2^2+k_2^2)$	$s_8s_{10} (g_2^2+k_2^2)$	$s_9s_{10} (g_2^2+k_2^2)$	$s_{10}^2(g_2^2k_2^2)+e_{10}^2$	
S11	$s_1s_{11} g_1g_2$	$s_2s_{11} g_1g_2$	$s_3s_{11}g_1g_2$	$s_4s_{11}g_1g_2$	$s_5s_{11} g_1g_2$	$s_6s_{11} g_1g_2$	$s_7s_{11} (g_2^2+k_2^2)$	$s_8s_{11} (g_2^2+k_2^2)$	$s_9s_{11} (g_2^2+k_2^2)$	$s_{10}s_{11} (g_2^2+k_2^2)$	$s_{11}^2(g_2^2k_2^2)+e_{11}^2$

Table 2. Expected covariance matrix for the nested factor model.

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**
S1	$s_1^2k_1^2 + g_1^2+e_1^2$										
S2	$s_1s_2k_1^2+g_1g_2$	$s_2^2k_1^2 + g_2^2+e_2^2$									
S3	$s_1s_3k_1^2+g_1g_3$	$s_3s_2k_1^2+g_3g_2$	$s_3^2k_1^2 + e_3^2+g_3^2$								
S4	$s_1s_4k_1^2+g_1g_4$	$s_4s_2k_1^2+g_4g_2$	$s_4s_3k_1^2+g_4g_3$	$s_4^2k_1^2 + e_4^2+g_4^2$							
S5	$s_1s_5k_1^2+g_1g_5$	$s_5s_2k_1^2+g_5g_2$	$s_5s_3k_1^2+g_5g_3$	$s_5s_4k_1^2+g_5g_4$	$s_5^2k_1^2 + e_5^2+g_5^2$						

Table 2. (Continued): Expected covariance matrix for the nested factor model.

S6	$s_1s_6k_1^2+g_1g_6$	$s_6s_2k_1^2+g_6g_2$	$s_6s_3k_1^2+g_6g_3$	$s_6s_4k_1^2+g_6g_4$	$s_6s_5k_1^2+g_6g_5$	$s_6^2k_1^2+e_6^2+g_6^2$					
S7	g_1g_7	g_2g_7	g_3g_7	g_4g_7	g_5g_7	g_6g_7	$s_7^2k_2^2+e_7^2+g_7^2$				
S8	g_1g_8	g_2g_8	g_3g_8	g_4g_8	g_5g_8	g_6g_8	$s_7s_8k_2^2+g_7g_8$	$s_8^2k_2^2+e_8^2+g_8^2$			
S9	g_1g_9	g_2g_9	g_3g_9	g_4g_9	g_5g_9	g_6g_9	$s_7s_9k_2^2+g_7g_9$	$s_8s_9k_2^2+g_8g_9$	$s_9^2k_2^2+e_9^2+g_9^2$		
S10	g_1g_{10}	g_2g_{10}	g_3g_{10}	g_4g_{10}	g_5g_{10}	g_6g_{10}	$s_7s_{10}k_2^2+g_7g_{10}$	$s_8s_{10}k_2^2+g_8g_{10}$	$s_{10}s_9k_2^2+g_{10}g_9$	$s_{10}^2k_2+e_{10}^2+g_{10}^2$	
S11	g_1g_{11}	g_2g_{11}	g_3g_{11}	g_4g_{11}	g_5g_{11}	g_6g_{11}	$s_7s_{11}k_2^2+g_7g_{11}$	$s_8s_{11}k_2^2+g_8g_{11}$	$s_{11}s_9k_2^2+g_{11}g_9$	$s_{10}s_{11}k_2^2+g_{10}g_{11}$	$s_{11}^2k_2^2+e_{11}^2+g_{11}^2$

Figures 8 and 9 show probabilities of exceeding the estimated distances between the pseudosolution residual vector $\boldsymbol{\varepsilon}$ and its corresponding SOFM "winning" unit center for different standard deviations of pseudosolution components and their changing averaged percentages of going beyond the given neighborhood intervals. Given probability distributions were computed by means of the presented method including the discussed factor model identification, SOFM and Monte Carlo techniques.

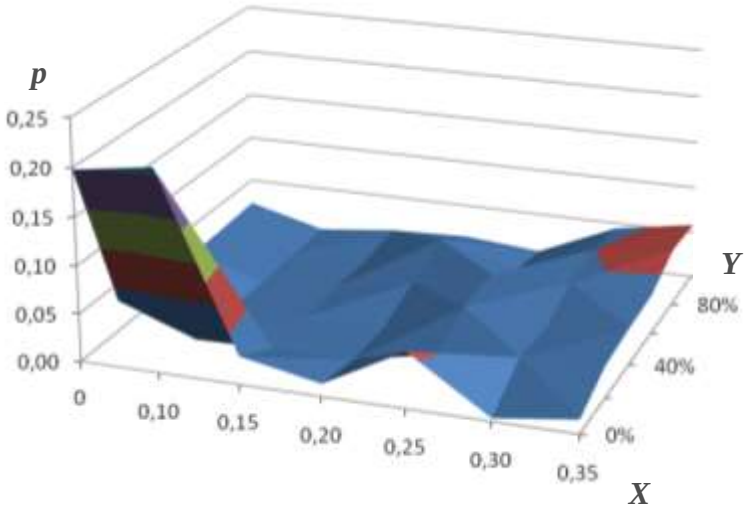

Figure 8. Hierarchical factor model: distribution of probabilities of exceeding the estimated distances between the pseudosolution residual vector and its corresponding SOFM "winning" unit center for different standard deviations of pseudosolution components (axis *X*) and their changing averaged percentages of going beyond the given neighborhood intervals (axis *Y*).

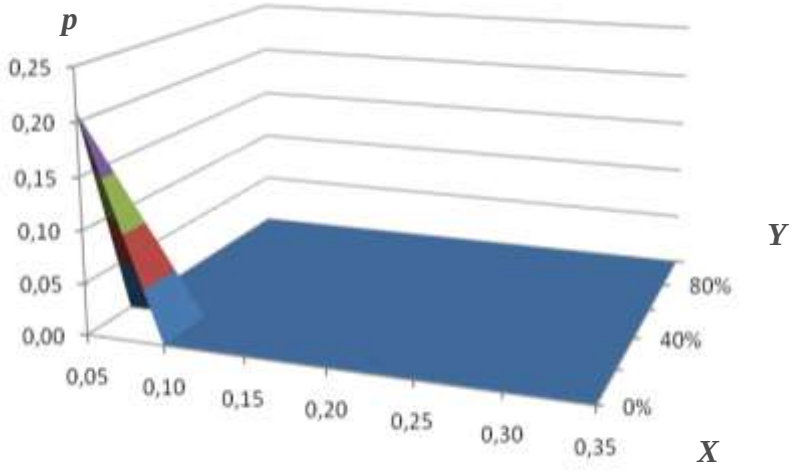

Figure 9. Nested factor model: distribution of probabilities of exceeding the estimated distances between the pseudosolution residual vector and its corresponding SOFM "winning" unit center for different standard deviations of pseudosolution components (axis *X*) and their changing averaged percentages of going beyond the given neighborhood intervals (axis *Y*).

Maximum likelihood ratios r_1 and r_2 equal to 0.05 and 0.1, correspondingly. Since $m_2=0.5$ and $\sigma_*=0.05$, probability $P(m_2 \le X \le 1)=\Phi(1)-\Phi(0.5)$ of the obtained mean deviation equals 0.5. Therefore, it may be concluded that the hierarchical model meets observation data significantly better that the nested model at the significance level 0.05.

5 Main results and conclusions

Further development of the new approach to goodness-of-fit factor model analysis, which is based on the capabilities of self-organizing feature maps and the Monte Carlo method, was proposed to avoid undesirable restrictions on both factor models and observation data. Advantages of this technique are:

- Arbitrary structural equation factor models without reciprocal causal relationships of variables leading to "feedback-loops" are acceptable
- No need to test multivariate normality of distributions of either observed variables or residual vector components
- It is possible to reveal the maximum likelihood combination of the obtained pseudosolution precision and the component-wise structure of significant deviations for the pseudosolution components
- Simple procedure of estimating type 2 statistical errors is available
- Higher reliability of obtained goodness-of-fit measures because of unrestrictedness of generated random samples of the pseudosolution components and the following unlimited goodness-of-fit estimation accuracy.

Acknowledgements. This work is supported by grant 14-06-00191 from the Russian Foundation for Basic Research.

References

[1] J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*, John Wiley and Sons, New York, 1986.

[2] T. Kohonen, Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, **43** (1982), 59-69. http://dx.doi.org/10.1007/bf00337288

[3] L. S. Kuravsky and S. N. Baranov, Development of the wavelet-based confirmatory factor analysis for monitoring of system factors - *In: Proc. 5th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, Edinburgh, United Kingdom, (2008), 818-834.

[4] L. S. Kuravsky, P. A. Marmalyuk, S. N. Baranov and N. I. Baranov, Wavelet-Based Confirmatory Factor Analysis: Monitoring of Damage Accumulation

Factors, *Applied Mathematical Sciences*, **9** (2015), no. 26, 1245-1263. http://dx.doi.org/10.12988/ams.2015.4121026

[5] L. S. Kuravsky, S. N. Baranov and P. A. Marmalyuk, Estimation of goodness-of-fit measures accompanying the identification of factor models - *In: Proc. 7th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies*, Stratford-upon-Avon, England, (2010).

[6] L. S. Kuravsky, P. A. Marmalyuk, S. N. Baranov, V. I. Alkhimov, G. A. Yuryev and S. V. Artyukhina, A New Technique for Testing Professional Skills and Competencies and Examples of its Practical Applications, *Applied Mathematical Sciences*, **9** (2015), no. 21, 1003-1026. http://dx.doi.org/10.12988/ams.2015.411899

[7] L. S. Kuravsky, P. A. Marmalyuk, G. A. Yuryev and P. N. Dumin, A Numerical Technique for the Identification of Discrete-State Continuous-Time Markov Models, *Applied Mathematical Sciences*, **9** (2015), no. 8, 379-391. http://dx.doi.org/10.12988/ams.2015.410882

[8] L. S. Kuravsky, P. A. Marmalyuk, G. A. Yuryev, P. N. Dumin and A. S. Panfilova, Probabilistic Modeling of a Testing Procedure, *Applied Mathematical Sciences*, **9** (2015), no. 82, 4053 - 4066. http://dx.doi.org/10.12988/ams.2015.53234

[9] M. C. Neale, L. R. Cardon, *Methodology for Genetic Studies of Twins and Families*, Dordrecht, The Netherlands, Kluwer Academic Publishers, 67, 1992. http://dx.doi.org/10.1007/978-94-015-8018-2

[10] D. V. Ushakov, *Intelligence: Structural and Dynamic Theory*, Institute of Psychology of Russian Academy of Sciences, Moscow, Russia, 2003, in Russian.

Received: July 28, 2015; Published: March 10, 2016